Measuring News Similarity Across Ten U.S. News Sites

Grant C. Atkins
Old Dominion University
Norfolk, Virginia
gatkins@cs.odu.edu

Alexander C. Nwala
Old Dominion University
Norfolk, Virginia
anwala@cs.odu.edu

Michele C. Weigle
Old Dominion University
Norfolk, Virginia
mweigle@cs.odu.edu

Michael L. Nelson
Old Dominion University
Norfolk, Virginia
mln@cs.odu.edu

ABSTRACT

News websites make editorial decisions about what stories to include on their website homepages and what stories to emphasize (e.g., large font size for main story). The emphasized stories on a news website are often highly similar to many other news websites (e.g, a terrorist event story). The selective emphasis of a top news story and the similarity of news across different news organizations are well-known phenomena but not well-measured. We provide a method for identifying the top news story for a select set of U.S.-based news websites and then quantify the similarity across them. To achieve this, we first developed a headline and link extractor that parses select websites, and then examined ten United States based news website homepages during a three month period, November 2016 to January 2017. Using archived copies, retrieved from the Internet Archive (IA), we discuss the methods and difficulties for parsing these websites, and how events such as a presidential election can lead news websites to alter their document representation just for these events. We use our parser to extract $k = 1, 3, 10$ maximum number of stories for each news site. Second, we used the cosine similarity measure to calculate news similarity at 8PM Eastern Time for each day in the three months. The similarity scores show a buildup (0.335) before Election Day, with a declining value (0.328) on Election Day, and an increase (0.354) after Election Day. Our method shows that we can effectively identity top stories and quantify news similarity.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**;

KEYWORDS

Web Archiving, Document Representation, Similarity

ACM Reference Format:
Grant C. Atkins, Alexander C. Nwala, Michele C. Weigle, and Michael L. Nelson. 2018. Measuring News Similarity Across Ten U.S. News Sites. In *Proceedings of International Conference on Digital Preservation (IPRES'18).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

IPRES'18, September 2018, Boston, Massachusets USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION & MOTIVATION

We are interested in mining archived news websites, specifically to measure similarity of news across different sites. This involves identifying what is considered news, how it is represented, and emphasized on a website. For example, on January 4, 2018 *foxnews.com*'s coverage of the *Fire and Fury* book [26] was not apparent on their website[1]; it was mid-way down the page with the headline "Trump demands publisher halt release of book that led to Bannon fallout." Their top story was "Freedom to drill: Trump dramatically expands offshore drilling, opens nearly all US coastal waters to explore oil and gas". At roughly the same time, *msnbc.com*'s top story was *Fire and Fury*, with the top story entitled "Trump: Bannon changed tune, called me a great man last night" and there were no stories about offshore drilling[2] "above the fold" (i.e., viewable in the top portion of the webpage without scrolling).

Most of the journalistic work that chronicles important events is represented in digital form as online news content. Consequently, there has been an increased interest in preserving online news. For example, the "Dodging the Memory Hole" initiative [21] emphasizes the importance of long-term preservation of online news content. We consider the preservation of online news important as well as the provision of tools[3] to study and perform analysis on archived news content. News organizations, like *The New York Times*, which recently started their own archiving initiatives [24], also consider the preservation of old pages to be of importance. As described by Hansen & Paul [16], many news websites, like *The Wall Street Journal*, do not have screen captures of when their websites officially launched. Web archives have become a key tool to go back in time and replay these websites.

Web archives can only offer what they have, requiring new tools [4] to preserve online content in multiple archives. Comparing archived news in aggregate also requires a comparison of their preserved date time. It is not expected for web archives to have continuous minute-by-minute preservations of web pages which is a tradeoff when compared to the live web.

To measure similarity in news websites, we retrieve the mementos (i.e., archived web pages) for the months of November 2016, December 2016, and January 2017 from the top-level page of 10 national news websites based in the United States. We extracted the

[1] http://web.archive.org/web/20180104210001/http://www.foxnews.com/
[2] http://web.archive.org/web/20180104203948/http://www.msnbc.com/
[3] http://www.pastpages.org/

headlines and URIs for the top-k (where $k = 1, 3, 10$) news stories as close as possible to 8PM Eastern time for all 10 sites.

We propose a new tool to parse select news sites HTML documents using CSS selectors to identify top stories and other top headlines. We considered RSS feeds for this task as they may offer top news stories, however, they are often provided in the order they are created, which may not reflect how an actual homepage for a news site might look. RSS feeds are also not always reliably or as frequently archived as news homepages.

Our tool aims to solve this issue by identifying the top stories for a specified news site. Additionally, our tool identifies "Hero stories," which are the most prominent top stories on news sites often emphasized with large font and central placement. A Hero story, where $k = 1$, is identified by position, font size, and image size (if one exists), depending on the layout of the news homepage. If our parser failed to extract the top news story we considered the next subsequent headline parsed as the Hero story.

We explore the use of the cosine similarity measure to calculate the similarity of a collection. Over a three month period we show that the count of stories used in the news similarity calculation directly influences the overall similarity scores for a given day. We show that using these similarity scores we can effectively identify events such as the 2016 presidential election, a national holiday like Thanksgiving, and other significant political events like the announcement of the Travel Ban [1] (Executive Order 13769).

Using the cosine similarity scores we identify the Travel Ban event to have the highest overall similarity score regardless of the value of k. We show in Section 4.1 that our similarity score can be used to identify overlapping stories for a given day outside of the election. We identify that similarity values peak after significant events start. For example, after Election Day and the Travel Ban announcement cosine similarity scores increase, indicating that there is a delay in synchronization of news. Our similarity scores show that we are able to identify a decline, from 0.417 to 0.343, in similarity after the U.S. election period has passed, which indicates news sites pursuing other stories.

2 RELATED WORK

There are many existing efforts already available in topic detection, news parsing, and building collections. Topic Detection and Tracking (TDT) [3], a DARPA-sponsored initiative, introduced and formalized the problem of determining if two new stories are about the same topic or highly similar. The ability to determine similarity enables many tasks such as clustering news. Lau et al. [19] developed a method to track emerging events on Twitter using a topic model based on time slices and a dynamic vocabulary. He et al. [17] modeled the problem of detecting bursts in news cycles by using physics concepts such as mass and velocity. They showed this approach was effective at accurately and efficiently detecting bursts for MeSH (Medical Subject Heading) terms. Similar to these approaches, we modeled news reporting of the November 2016 U.S. election cycle and showed the similarity measure we defined is effective in detecting news similarity and bursts of news synchronization.

We used archived webpages, or mementos, from the Internet Archive in our analysis. As demonstrated by Brunelle et al. [7] and Berlin [6], not all mementos render correctly upon playback. This affects our method to determine the top stories on a news website. In Section 4.2, we discuss the representations of some of the news websites whose mementos from November 2016 are not rendered consistently.

Klein and Broadwell [18] presented analyses on both television news and social media collections showing spikes in attention for continuous news stories and the quick drops that follow afterwards.

Other efforts related to news parsing can be seen in studies conducted by various news organizations. The *New York Times* conducted a study [25] that showed how three different news organizations (*Fox News*, *msnbc.com*, and *CNN*) covered the indictment of Paul Manafort and Rick Gates by the special counsel Robert Mueller. They showed that *msnbc.com* and *CNN* devoted more airtime to the coverage of this news. Similarly, *FiveThirtyEight* [11] showed that the ideological leanings affected when/what news networks covered about the "Trump-Russia" story. Even though our research identifies common interests in news reports during the 2016 U.S. election cycle, we do not infer the reasons behind why the news organizations cover various topics. However, our analysis could help a domain expert conduct such a study.

Similarity in news has also been studied in the context of identifying media manipulation on social media. Wolley et al. [27] studied propaganda bot activities during the 2016 U.S. general election and showed how these bots propagate similar content to boost the prominence of a political candidate in an attempt to "manufacture consensus". Roger Sollenberger [22] showed a similar activity among fringe media organizations by illustrating how they publish similar content and interlink their web pages in an attempt to increase their search engine rankings. Similarly, Faris et al. [13] studied the media landscape of the 2016 U.S. general election in order to identify how partisan news organizations covered Donald Trump and Hillary Clinton. They showed that the media overwhelmingly covered Trump more than Clinton. These research efforts show that a high degree of similarity between news content is not always due to an organic increase in interest surrounding a topic, but it can be manufactured. Our research identifies similarity over time, and since we control the news sources we sample URIs from, we do not focus on assessing if any high degree similarity in news content is inorganic.

Many other research efforts that combine topic detection and news parsing are tied toward collection building. Nwala et al. [20] introduced the Local Memory Project, which provided a suite of tools for archiving, building, and exploring collections from local news stories. Hamborg et al. [15] introduced NewsBird, a news aggregation system that attempts to limit bias in news collections by balancing multiple different perspectives on international news topics. Other efforts related to collection building use focused crawlers in order to build collections about specific topics. At the center of focused crawling is the means to determine if an incoming URI belongs in a collection. Some systems [5, 9] used classifiers to determine if a URI belongs in a collection, others [12] have used a similarity measure. Even though this research is not focused on collection building, it is relevant to collection building efforts since the news similarity metric can serve as a filter in the extraction of URIs that share a common topic.

3 METHODOLOGY

Selecting news homepages outside of a relatively close timeframe can greatly alter the evaluation of the similarity. Suppose a user reads the homepage of news website in the morning, do they expect to see the same homepage by the end of the day? Although it is important that each of these news websites are archived frequently to catch the fast changing news, for our purpose we needed to establish a time from which to collect a single memento where each news homepage was archived and then evaluate upon the set of mementos collected at this time. This section is organized as follows. First, we explain why our methodology is only possible due to the preservation of news sites in web archives. Second, we provide an explanation on the chosen news websites and memento collection time. Third, we explain the process to extract the top story and subsequent headline titles and links from each of these homepages. Finally, we describe the similarity measure used to calculate the similarity for a given day over the course of the 3 selected months.

3.1 Mining Archived News

There is an increased interest in archiving news articles and homepages. When these news pages become archived, they can be used for more than just replay. For example, archived news pages can be used to study document design, significant events, political biases, editorial decisions, video playback, etc.

The methodology proposed in the following sections is not practical for live webpages mainly due to the following reasons:

- Parsing news homepages based on CSS selectors requires document representations to be static - unchanging over time. We can build static CSS parsers because we can study archived copies for these pages. This is not possible on the live web due to the fact that sites often change their document representations or CSS naming conventions, which will result in the parsing of zero stories for future time periods.
- Live web pages can introduce noise into similarity calculations. In Section 3.2 we discuss the reasons why we chose not to select specific news sites due to paywalls. Duplicate pages such as paywalls [14], login pages, or empty documents, introduce noise when calculating similarity between documents which is not only limited to news content. Although archives may also contain the aforementioned pages, it becomes a simpler task using archives to find such pages whereas on the live web it is difficult to differentiate between a news article or a news paywall page.

The challenges mentioned in this section and the document representation changes mentioned in Section 4.2 show that web archives are a key tool in calculating similarity for collections of news pages. The use of preserved web pages allows the collection and analysis of news webpages in a controlled way.

3.2 Memento selection

Table 1 outlines the ten U.S. based newspaper and TV organizations we considered for our analysis.

The *Wall Street Journal* (*WSJ*) was also considered for this project, however we discovered a majority of its stories are behind a paywall requiring users to subscribe to their website to view more than a snippet for a story. Investigations into archived paywalls [14, 23] show that news sites are increasingly putting content behind paywalls. Therefore, we excluded *WSJ* and other sites that only provided snippet stories.

Table 1: Memento counts for news site homepages for November 2016, December 2016, and January 2017.

Site	Nov. 2016	Dec. 2016	Jan. 2017
washingtonpost.com	3560	3950	3836
foxnews.com	1250	1459	1313
abcnews.go.com	708	874	1193
nytimes.com	3177	3936	3809
usatoday.com	1405	1773	1546
cbsnews.com	816	741	805
chicagotribune.com	470	609	565
nbcnews.com	818	1164	1048
latimes.com	950	1188	989
npr.org	2208	3016	2371

Another site considered for this project was *msnbc.com*. Although the archival rate for this site in November 2016 was high, most of the stories shown on the homepage were actually multimedia and would lead to small snippet summaries or no summary. This website's **class** attribute naming conventions did not differentiate between the textual stories and the multimedia stories and may cause our parser to introduce false positives. Therefore *msnbc.com* was also excluded.

We omitted *CNN* because of known issues [6] with replaying their mementos from the Internet Archive. Unfortunately, these issues began in November 2016, the start of our target timeframe.

News sites like *CNN*, *WSJ*, and *msnbc.com*,are not the only sites that have mementos that are afflicted with problems. When archived sites render their content, stylesheets and JavaScript dependencies may be missing, making it difficult to determine the order and importance of content due to the lack of structure incurred by the missing dependencies. News sites with paywalled content may resolve [14] in time to actual content in the Internet Archive, but this could also lead to the selection of a different page at another point in time. These are a few of the hazards of working with and relying on the content of web archives.

We used the Internet Archive's Memento API to collect the TimeMaps (TM), which are lists consisting of the URIs for archived copies (mementos) of webpages, for each of the news websites in Table 1. From these TMs we selected all of the mementos from November 2016 to January 2017. We selected this date range because we acknowledged that a significant event, the United States presidential election, occurs along with three other national holidays: Veterans Day (November 11), Thanksgiving (November 24), and Christmas (December 25), which provided the opportunity to identify these events.

When looking at the archival time for the mementos for the month of November, we found that most of the websites were archived at approximately 1AM GMT, or 8PM Eastern Time, as shown in Figure 1. For the three months, 8 PM Eastern Time had 335 mementos archived at this time while 7 AM Eastern Time had lowest average number of archived mementos at 154 mementos.

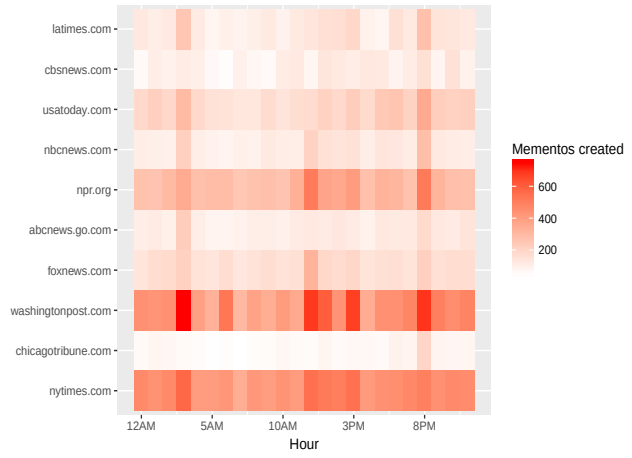


Figure 1: Memento creation times by hour, Eastern Time, for the 3 selected months. This shows that 8PM has the highest archival rate.

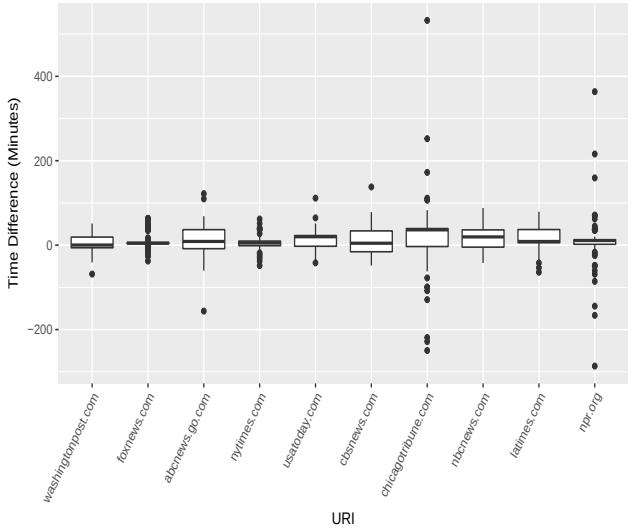


Figure 2: Time difference of Mementos collected for each news website from the request time of 8 PM Eastern Time

We therefore chose 8 PM Eastern Time to ensure we had a higher chance to obtain mementos closer to the same archived time. We subsequently collected the mementos from the TM at 1AM GMT (8 PM Eastern Time) from 2016-11-01 to 2017-01-31 using the Internet Archive's Memento API.

We found that the distribution of minutes from the requested memento time is on average within 100 minutes of the original request, as shown in Figure 2. Figure 2 shows the distribution of the distance between the datetime of the URI-M (Memento URI) and the requested time of 8 PM Eastern Time in minutes. The boxes for each news site represent the overall distribution between their mementos. A smaller box range shows that a large amount of mementos were retrieved within a close time range. For example, *foxnews.com* has the smallest box in Figure 2 showing that most of the mementos in the Internet Archive were archived at almost the same time every day. While *npr.org* also has a small box, there is a greater number of outlier mementos, denoted by dots, that were not tightly archived around the majority time of 8 PM Eastern Time.

Figure 2 shows that *chicagotribune.com* and *npr.org* have multiple mementos beyond the 100 minute average range, showing that these websites may not be as popular for news and are therefore not archived as frequently. When comparing the memento counts shown in Table 1, washingtonpost.com has over 3000 mementos for each month, while *chicagotribune.com* has 400 to 600 mementos per month. This is an indication of the popularity of these sites, showing that less well known websites are less likely to be archived as frequently.

Many of the mementos requested at 8 PM Eastern Time returned mementos close to the request time of 8 PM Eastern Time. However, even if a site was archived thousands of times a month it does not mean that there would be a memento for every single hour or specified request time as indicated by *npr.org* in Figure 2. On average the time difference from the requested URI-M and the actual URI-M was 15.2 minutes, while the earliest memento collected was

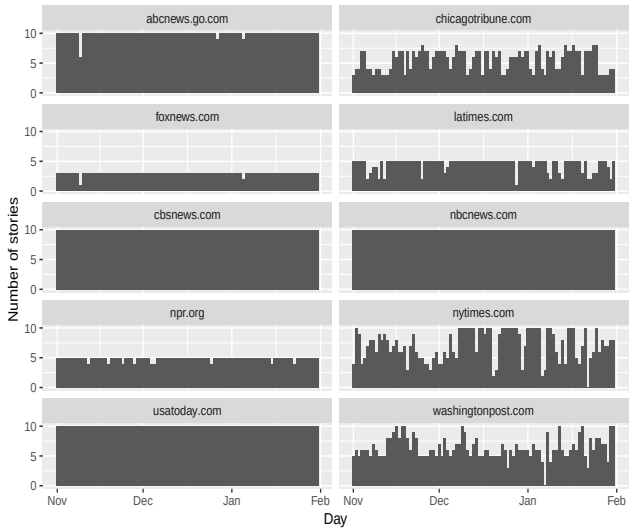
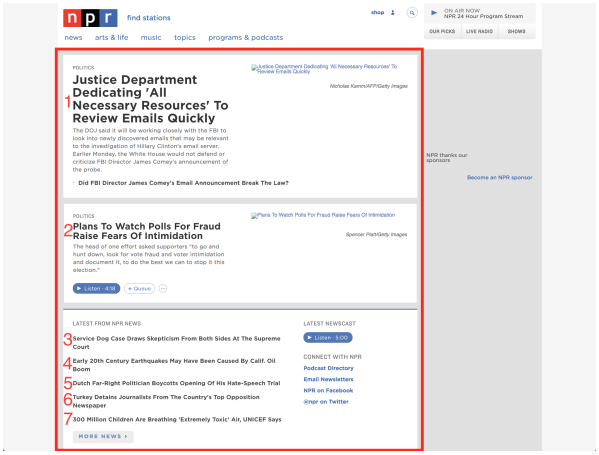


Figure 3: Total number of stories for each website, with a limit of k = 10, for each of the days from November 2016 to January 2017.

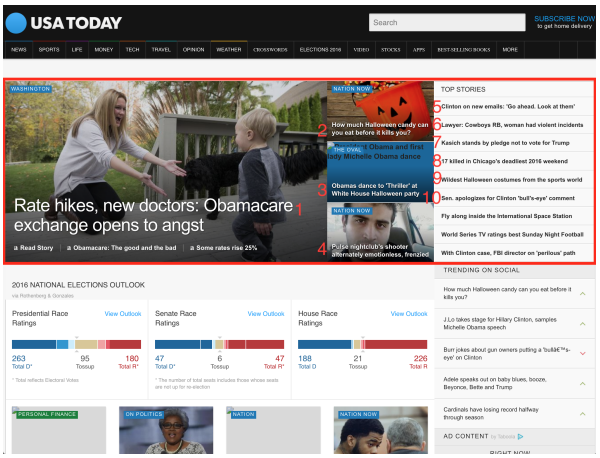
286.5 minutes before the requested time and the latest memento collected was 532.4 minutes after the requested time.

3.3 Homepage parsing

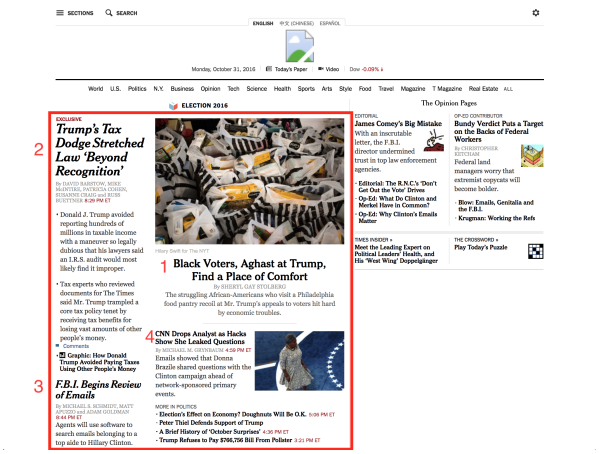
Parsing titles and retrieving all the links from an HTML document is a relatively simple task. This can be achieved by searching an HTML document for all the **<a>** elements that provide attributes, such as **href** or **src**. However, news websites may often contain hundreds of links either to recommendations, opinion stories, or any other category labeled by the news website. We use the term



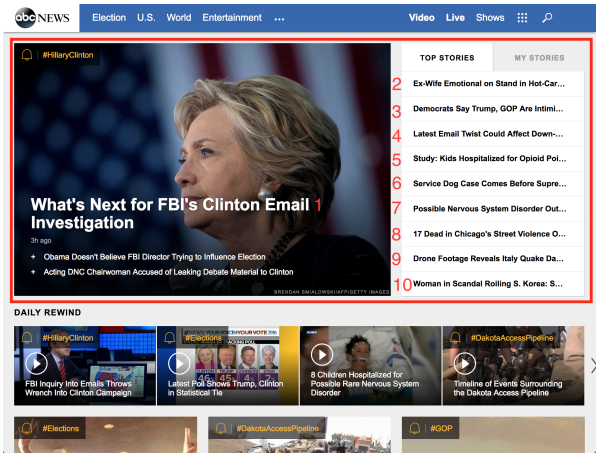
(a) NPR homepage



(b) USA Today homepage



(c) New York Times homepage



(d) ABC News homepage

Figure 4: Three mementos of news sites taken from November 1, 2016 1AM GMT (October 31, 2016 8PM Eastern Time). The number indicate the order of the stories processed and highlights the news site stories selected using CSS selectors. Number 1 indicates the Hero story for each news site. This day had cosine similarity score of 0.193 for $k = 3$ stories from each news site.

Hero stories in this paper to describe a story where the text is exaggerated across the top of a webpage or there is giant image indicating the main story at the current time on a website. For example, "Rate hikes, new doctors: Obamacare exchange opens to angst" as shown in Figure 4b. Our target stories were centered around the hero stories, where $k = 1$, or the main headlines. This often appears at the top of a news website and is usually presented when the page finishes loading without the need to scroll down.

To accomplish the task of custom parsing, we created a parser[4] utilizing Cascading Style Sheet (CSS) selectors, which are used to target HTML elements on web resources [8]. Our parser utilizes the Python package Beautiful Soup[5] to access CSS selectors inside HTML documents. Although CSS selectors are often used to apply formatting for a webpage, they also have applications to retrieve an element on a webpage. Therefore, we can use CSS selectors to filter elements based on their attributes, retrieve the attributes inside the elements, retrieve the children of elements of a specified selector, and retrieve the text based on a specified selector. CSS selectors prove to be fairly successful when testing against news websites, but when websites depend on HTML **iframes** or JavaScript-injected HTML, CSS selectors cannot be used to select content before it is loaded.

Identifying the stories to be considered as the Hero stories was done empirically depending on the layout of the webpage. For example, if the stories seemed to appear in a single column on the webpage (Figure 4a), the Hero story was usually denoted by the top story on the webpage that contained an enlarged image or enlarged text. Other websites try to emulate a newspaper's format by applying multiple columns in a single view, showing different categories of news available on their website. We found that when

[4]https://github.com/oduwsdl/top-news-selectors
[5]https://pypi.python.org/pypi/beautifulsoup4

choosing the Hero story for this type of format the main stories were often in the central column (Figure 4c), while the far right column stories were often opinion based or generated in real time.

Many of the websites explored in this paper self-identify the top stories presented in their website's HTML representation. For example, USA Today in Figure 4b shows the section our parser selects and the Hero story is identified by the selector "a.hfwmm-primary-hed-link". There is only one occurrence of this selector in the document. News sites that did not label their content with an obvious CSS selector were parsed by taking the top stories from the leftmost and center stories.

After identifying all of the stories for the new sites using CSS selectors we performed requests to the Internet Archives Memento API [10] to retrieve the content of each of the stories as provided from each memento. Figure 3 indicates the number of stories we identified for each news site for over the course of three months. We found that there were 25 HTTP 404 response codes, identified as archived paywalls, and two "infinite" redirection loops from the unique story URIs collected. For all documents that returned a 200 HTTP response code, we processed their HTML by extracting the text content through boilerplate removal with Python-Boilerplate[6] which we have found to consistently provide the best results [2].

3.4 Similarity Metrics

The collection similarity score, $s \in [0, 1]$, is a single value which quantifies the degree of similarity within the documents in a collection. A similarity score of 0 means all documents in the collection have no vocabulary in common, while a similarity score of 1 indicates maximum similarity (duplicate content). Given a collection of URIs, C, ($|C| = n$), the collection similarity score was calculated as follows:

(1) **Representation:** All the documents in the collection were represented as vectors. Specifically, the vector representation is where each document d_i in the collection C is represented as a vector of TF-IDF values.
(2) **Pairwise similarity:** The pairwise similarity of all the documents were calculated to populate a similarity matrix, $\boldsymbol{D} \in \mathbb{R}^{n \times n}$. Given a pair of documents, d_i and d_j in vector space, the similarity between the documents was calculated using the cosine similarity metric.
(3) **Collection similarity score:** Given an *all-ones matrix*, $\boldsymbol{O} \in \mathbb{R}^{n \times n}$, and a square matrix, $\boldsymbol{N} \in \mathbb{R}^{n \times n}$, with zeros on the main diagonal and ones everywhere else we can calculate the collection similarity s. For example if $\boldsymbol{N} \in \mathbb{R}^{3 \times 3}$,

$$\boldsymbol{N} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

The collection similarity, s would be calculated as follows: $s = \frac{||\boldsymbol{N}.\boldsymbol{D}||_F}{||\boldsymbol{N}.\boldsymbol{O}||_F}$, where $||\boldsymbol{A}||_F$ is the Frobenius norm:

$$||\boldsymbol{A}||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{i,j}|^2}$$

4 EXPERIMENT RESULTS

Using the mementos collected for the ten news sites we used our parser to extract each of the Hero stories and the subsequent headlines providing us with the text and the URI. For the three months we extracted a total of 2349 unique memento URIs[7]. We found that some websites used their CSS selectors for identifying structure and order priority of content. For example, *chicagotribune.com* had a Hero selector named "h2.trb_outfit_primaryItem_article_title" and a selector for identifying a section for leading headlines named ".trb_outfit_list_headline_a". Some of the websites, for example cbsnews.com, did not self identify where their headlines should cutoff. This would for some days lead to upwards of identifying twenty stories while other news sites would identify in a range of three to ten stories. An example of this is shown in Figure 4, where *USA Today* has at least ten top stories found while we identify seven and four stories for *NPR* and *New York Times*, respectively. Due to this reason, we decided to evaluate similarity limiting the number stories to a maximum of $k = 1, 3, 10$ for each news site. This means, when $k = 1$ we would have a maximum of ten stories, one story from each news site homepage. When $k = 3$ we would have a maximum of thirty stories, a maximum of three stories from each news site homepage. When $k = 10$ we would have a maximum of one hundred stories, a maximum of ten stories from each news site homepage. We discovered that there were occasions when accessing the Internet Archive Memento API, in which some of the stories requested either resulted in 404, 301, to many redirects, response codes. Therefore we excluded such URIs from the similarity calculation.

The position of stories in the HTML representations of the news site determines the order in which stories are determined for relevance, i.e. the third story found is story number three when $k = 3$. For the three mementos shown in Figure 4, there are stories that by just looking at their titles we can determine there is a similarity between them. For example, *NPR* recognizes their Hero story as "Justice Department Dedicating 'All Necessary Resources' To Review Emails Quickly" while *USA Today* and *New York Times* recognize their 5th and 3rd positioned top stories similar to *NPR*'s Hero story. For this day, if we take only $k = 1$ story for each news site, the similarity would be affected due to the recognition of this story's importance not being widely recognized as the Hero story for every news site. However, if we move on to $k = 3$ then the *New York Times* will have included their coverage of this story but also including two other stories that may not have a high similarity.

When identifying significant events we observed that political events had the most influence on the similarity scores. Shown in Figure 5, November 8, 2018 Election Day is an easily identifiable sequence of events but the peak of the similarity for this event occurs days after Election Day. When the executive order started on January 28, 2017 this period also shows that there is a buildup of events in similarity, but the peak of similarity scores is after the start date. This signifies that there is a delay in synchronization among news websites until the events have reached a critical point.

[6] https://github.com/misja/python-boilerpipe
[7] We make our dataset publicly available at: https://github.com/grantat/news-similarity

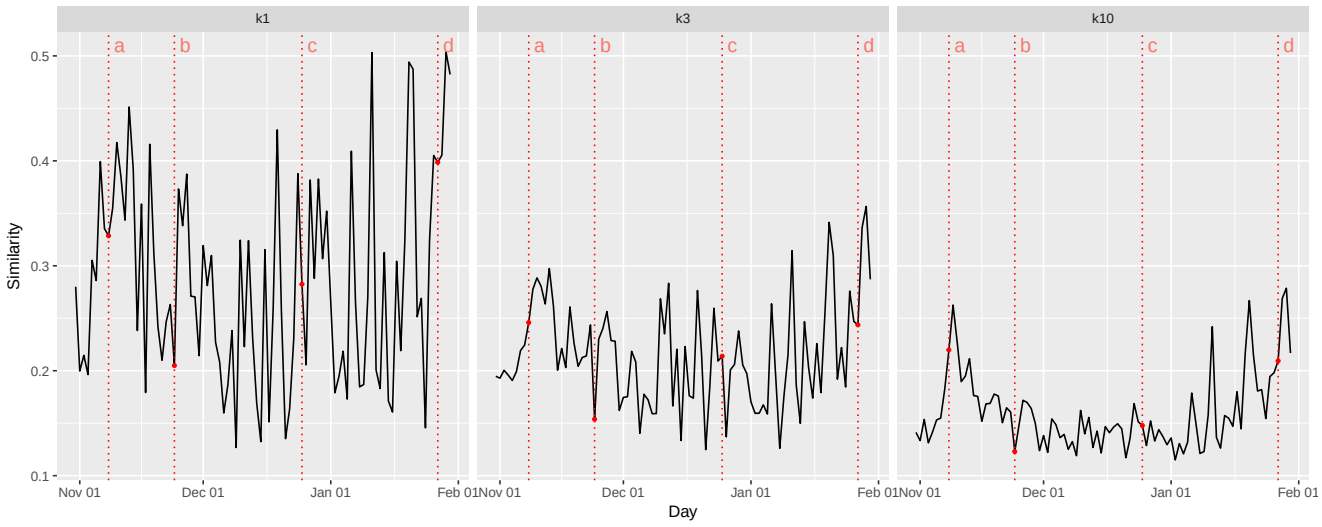


Figure 5: Cosine similarity scores for a given each day where $k = 1, 3, 10$ are the maximum number of stories for each news organization. Each graph is labeled with points of well known events: (a) Election Day (November 8, 2016), (b) Thanksgiving Day (November 24, 2016), (c) Christmas Day (December 25, 2016), and (d) Travel Ban (Executive Order 13769) comes into effect (January 27, 2017).

Figure 5 shows the cosine similarity scores, when $k = 1, 3, 10$, from November 2016 to January 2017 highlighting four different well known events. For each different k value we found the order of similarity between the four outlined events from highest to lowest was: Executive Order 13769 start date, Election Day, Christmas day, and Thanksgiving day. National holidays, such as Thanksgiving and Christmas, had relatively low similarity scores when $k = 10$. Specifically, Thanksgiving day was the 8th lowest similarity score of the 92 days compared. This shows that there is wider variance of stories and a lower synchronicity across the ten U.S. sites on national holidays.

4.1 Top-k Stories

When we limited the number of stories to the top $k = 1$ stories, we were essentially taking the Hero story from each of the news websites. This meant that we had a maximum of 10 stories for each of the days tested. We observed that when we limit the number of stories to $k = 1$ that cosine similarity becomes high for significant events but also has very high variance among days. Figure 5 shows that for November 8, 2016 there is a significant event that occurred preempted by a buildup of news similarity. Within a few days of the event begins to drop in cosine score indicating the story has either become less relevant or there is just less synchronization among news sites. Cosine score peaks during the end of January 2017 also preempted by a buildup of news similarity. During November 24, Thanksgiving day, there is a slight rise in similarity, however it is shown that the importance of this event is negligible compared to other stories that occurred around this day.

For $k = 3$ stories, Figure 5 shows a decline in similarity due to the introduction of more stories. The lowest similarity was seen when we took a maximum of $k = 10$ stories. This shows that as the number of stories increased the overall confidence that stories were similar decreased.

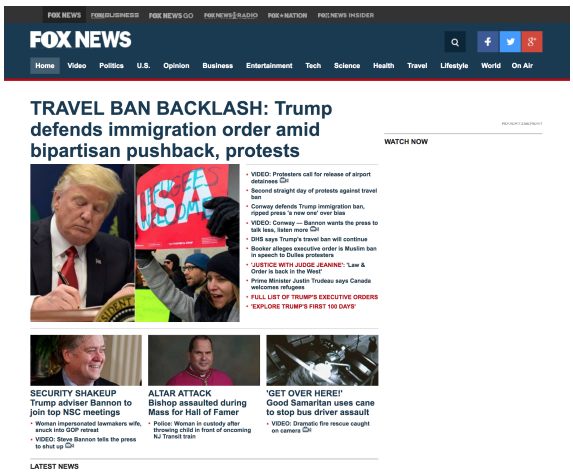
Table 2: Similarity score metrics for top-k stories for the 3 selected months.

Measure	Top-k	Min	Mean	Max
Cosine	1	0.1268	0.2858	0.5037
	3	0.1248	0.2160	0.3566
	10	0.1150	0.1608	0.2786

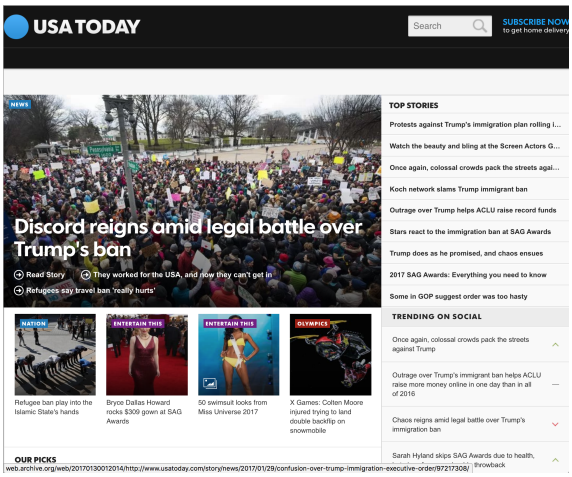
The Election Day has a clear buildup of similarity scores. Other days such as November 11, Veterans Day, also had a high similarity. Figure 6 shows there is a Hero story, related to the Travel Ban, shared across three different news sites. This shows that using these similarity metrics that we are able to recognize these synchronous events.

4.2 Election Day Influences

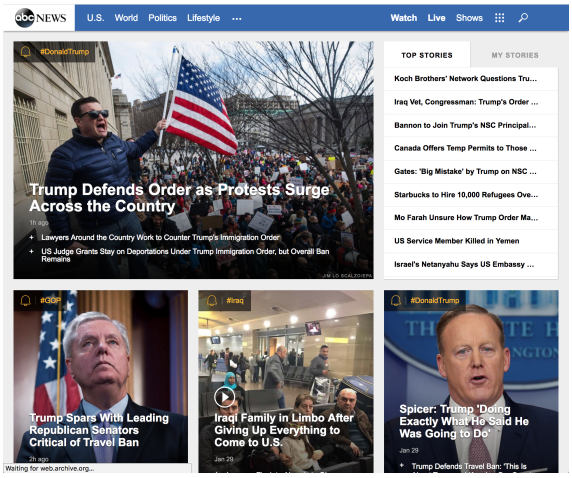
As shown in Figure 5, the election is a noticeable event due to high similarity scores. However these news sites also make this event noticeable by providing a new layout. When we first constructed our parser we noticed that using only a single set of CSS selectors proved to be ineffective when trying to parse headlines during the election. Between November 7th - 11th, we found that five of the ten news sites altered their document representation just for the United States presidential election. When these sites update their document representation some of them completely changed the naming conventions of CSS selectors but some of them still kept their previous selectors with this updated document. We therefore prioritized Election Day selectors over the default selectors.



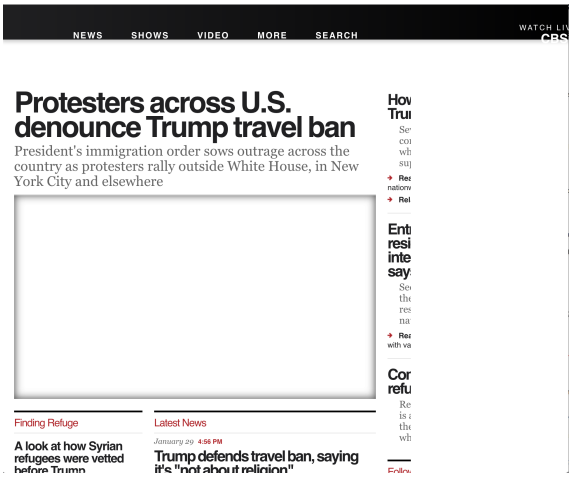
(a) Fox News homepage



(b) USA Today homepage



(c) ABC News homepage



(d) CBS News homepage

Figure 6: Four mementos of news sites taken from January 29, 2017 8PM Eastern Time. This day had the highest cosine similarity score for each maximum k value, where $k = 1$ was 0.504, $k = 3$ was 0.357, and $k = 10$ was 0.279. These mementos do not discuss a national holiday and are a separate event from the presidential election but showed a high similarity because they address the same Hero story, $k = 1$. The Hero story for each news site was related to the "Trump travel ban."

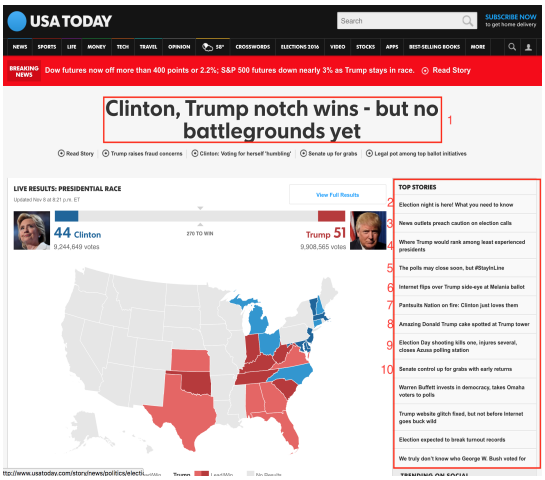
This also included the introduction of new naming conventions for HTML element **class** and **id** attributes. To address this we added multiple selectors for these websites, some for both the Hero stories and also the subsequent headlines.

These updates to their site layout meant that these sites recognized the significance of this event and chose to provide a new document representations to emphasize the importance of this event. Figure 7 shows these changes and that many sites often included a United States map to track the progress of a state's electoral college votes for either presidential candidate. After November 11th these sites returned to their original HTML representation.

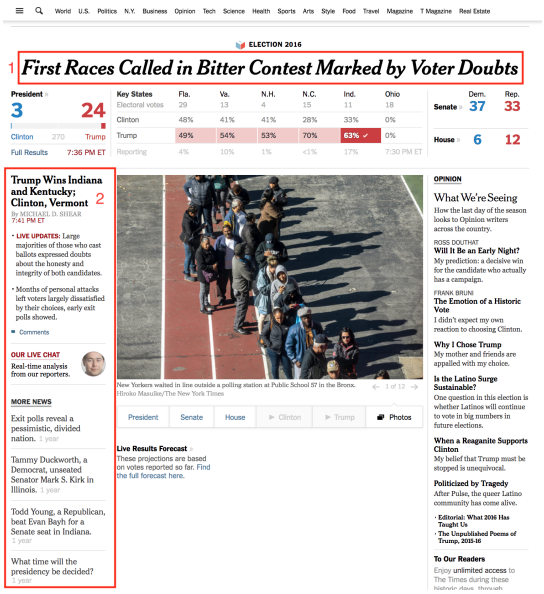
5 FUTURE WORK

For our experiment we only took news homepage mementos for a period of three months, we would like to extend the range to see how similarity changes. We selected a single time in day to be used repeatedly across our time period, much like Klein and Broadwell [18]. We would like to perform continuous observation with shorter intervals on a news site homepage and see how the document changes as well as the similarity to other news sites changes. We realized that CSS selectors are an easily accessible format that allows us to select stories but as news websites update their document representations and change class naming conventions we also have to update the range of selectors our parser uses.

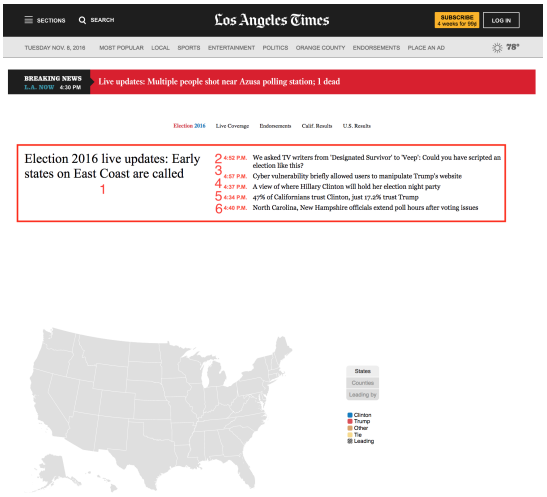
We found that a majority of titles for news homepage stories were actually shortened or summarized titles of the actual stories



(a) USA Today homepage



(b) New York Times homepage



(c) Los Angeles Times homepage

Figure 7: News sites format updated to track Election Day (November 8, 2016) progress for candidates: (a) USA Today introduces a United States map tracking candidate progress altering the naming conventions of their CSS selectors, (b) New York Times introduces a table of percentages of electoral votes, and (c) Los Angeles Times also introduces a world map to track candidate progress.

they reference. We wish to analyze this further and see how the similarity of titles presented on the homepage differs from what is actually considered the true title of a story.

6 CONCLUSIONS

The preservation of news is a valuable part of saving the memory of important historical events. Archived news pages provide a valuable opportunity for studying and analyzing events in a manner not possible on the live web. We provide tools to aid the analysis of archived news webpages in this work by introducing tools for parsing select HTML news. These tools allow for parsing select HTML news sites for Hero and headline stories using CSS selectors. We explored measuring similarity for ten U.S. sites using the cosine similarity measure. We also discuss how news sites may alter their document representations for significant events such as a presidential election.

We define a method of mining web archives, specifically mining archived news. We identify potential hazards when choosing mementos and describe the choices for which archived news sources are applicable for experimenting upon.

Our experiments for a three month period have shown that as the number of stories increase the overall similarity decreased. Using the calculated cosine scores we identify a decline, from 0.417 to 0.343, in similarity after the U.S. election period has passed, which indicates news sites pursuing other stories and decreased synchronization. Our results show these measurements can be used to identify synchronous stories outside of related national events for a given day which allowed us to identify the rise and decline of coverage using similarity.

ACKNOWLEDGMENTS

This work supported in part by NSF III 1526700 and IMLS LG-71-15-0077-15.

REFERENCES

[1] 2017. Executive Order Protecting the Nation from Foreign Terrorist Entry into the United States (Executive Order 13769). https://www.whitehouse.gov/presidential-actions/executive-order-protecting-nation-foreign-terrorist-entry-united-states/. (2017).

[2] Alexander Nwala. 2017. A survey of 5 boilerplate removal methods. http://ws-dl.blogspot.com/2017/03/2017-03-20-survey-of-5-boilerplate.html. (2017).

[3] James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study final report. (1998).

[4] Mohamed Aturban, Mat Kelly, Sawood Alam, John Berlin, Michael Nelson, and Michele Weigle. 2018. ArchiveNow: Simplified, Extensible, Multi-Archive Preservation. (2018).

[5] Donna Bergmark. 2002. Collection synthesis. In *Joint Conference on Digital Libraries (JCDL 2002)*. 253–262.

[6] John Berlin. 2017. 2017-01-20: CNN.com has been unarchivable since November 1st, 2016. http://ws-dl.blogspot.com/2017/01/2017-01-20-cncom-has-been-unarchivable.html. (Jan. 2017).

[7] Justin F Brunelle, Mat Kelly, Hany SalahEldeen, Michele C Weigle, and Michael L Nelson. 2015. Not all mementos are created equal: Measuring the impact of missing resources. *International Journal on Digital Libraries* 16, 3-4 (2015), 283–301.

[8] Tantek Celik, Elika J. Etemad, Daniel Glazman, Ian Hickson, Peter Linss, and John Williams. 2011. Selectors Level 3. (Sept. 2011). Retrieved January 15, 2017 from https://www.w3.org/TR/css3-selectors/

[9] Soumen Chakrabarti, Martin Van den Berg, and Byron Dom. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer networks* 31, 11 (1999), 1623–1640.

[10] H. Van de Sompel, M. Nelson, and R. Sanderson. 2013. *HTTP Framework for Time-Based Access to Resource States – Memento.* RFC 7089. RFC Editor.

[11] Dhrumil Mehta. 2017. All The Cable News Networks Are Covering The 'Russia Story' - Just Not The Same One. https://fivethirtyeight.com/features/all-the-cable-news-networks-are-covering-the-russia-story-just-not-the-same-one/. (2017).

[12] Mohamed MG Farag, Sunshin Lee, and Edward A Fox. 2018. Focused crawler for events. *International Journal on Digital Libraries (IJDL)* 19, 1 (2018), 1–19.

[13] Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. 2017. Partisanship, Propaganda, and Disinformation: Online Media and the 2016 US Presidential Election. (2017).

[14] Grant Atkins. 2018. Paywalls in the Internet Archive. http://ws-dl.blogspot.com/2018/03/2018-03-15-paywalls-in-internet-archive.html. (2018).

[15] F. Hamborg, N. Meuschke, and B. Gipp. 2017. Matrix-Based News Aggregation: Exploring Different News Perspectives. In *Joint Conference on Digital Libraries (JCDL 2017)*. IEEE, 1–10.

[16] Kathleen A. Hansen and Nora Paul. 2017. *Future-proofing the news : preserving the first draft of history.* Rowman & Littlefield, Lanham.

[17] Dan He and D Stott Parker. 2010. Topic dynamics: an alternative model of bursts in streams of topics. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 443–452.

[18] Martin Klein and Peter Broadwell. 2015. Analyzing News Events in Non-Traditional Digital Library Collections. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '15)*. 191–194.

[19] Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line Trend Analysis with Topic Models: # twitter Trends Detection Topic Model Online.. In *COLING*. 1519–1534.

[20] Alexander C Nwala, Michele C Weigle, Adam B Ziegler, Anastasia Aizman, and Michael L Nelson. 2017. Local Memory Project: Providing Tools to Build Collections of Stories for Local Events from Local Sources. In *Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on*. IEEE, 1–10.

[21] Reynolds Journalism Institute. 2017. Dodging the Memory Hole 2017. https://www.rjionline.org/events/dodging-the-memory-hole-2017. (2017).

[22] Roger Sollenberger. 2017. How the Trump-Russia Data Machine Games Google to Fool Americans. https://www.pastemagazine.com/articles/2017/06/how-the-trump-russia-data-machine-games-google-to.html. (2017).

[23] David Rosenthal. 2017. Talk at Spring 2013 CNI. https://blog.dshr.org/2013/04/talk-at-spring-2013-cni.html. (December 2017).

[24] Shan Wang. 2018. Here's how The New York Times is trying to preserve millions of old pages the way they were originally published. http://www.niemanlab.org/2018/04/. (2018).

[25] Taylor Adams, Jessia Ma and Stuart A. Thompson. 2017. Trump Loves 'Fox & Friends.' Here's Why. https://www.nytimes.com/interactive/2017/11/01/opinion/How-Fox-News-Covered-the-Manafort-Indictment.html. (2017).

[26] Michael Wolff. 2018. *Fire and Fury: Inside the Trump White House.* Henry Holt and Co.

[27] Samuel C Woolley and Douglas R Guilbeault. 2017. Computational Propaganda in the United States of America: Manufacturing Consensus Online. *Computational Propgaganda Research Project* (2017), 22.